Chapter 5
ANALOG MODULATION

When assessing a communications link for its susceptibility to jamming, among the attributes which must be considered is the type of signal modulation. Modulation schemes are divided into two major classifications: analog and digital. An analog signal is one which can take on an infinite number of values (consider a sinusoid signal) while a digital signal can have only a discrete number (for example a square wave only takes on two values). These two classes are further subdivided into several different methods. Each method has characteristics which are unique and others which it shares with the other methods. It is these characteristics which must be understood and exploited in order to achieve success in denying or degrading the link signal at the receiving site(s) with proper jammer waveforms. In this chapter the signal structures of the different methods of analog modulation will be investigated. In Chapter 6 we will analyze signals with digital modulation.

Modulation is a process which modifies over time a carrier signal in some specific manner. A carrier signal c(t) is a sinusoid at a specified, constant frequency f_c. Its amplitude A_c is likewise constant and known, so that a suitable mathematical model to represent the carrier signal is (see Section 2.3)

$$c(t) = A_c \cos(2\pi f_c t + \theta), \qquad \textbf{(5-1)}$$

where θ may or may not be known but is constant. Since the carrier signal is normally taken as the reference, θ is usually taken to be zero. No matter how the carrier signal is modified, the result must conform to the sinusoidal general form (given in Section 2.4). A modulated signal s(t) must therefore follow the model

$$s(t) = A(t) \cos (2\pi f_c t + \theta(t)). \quad \textbf{(5-2)}$$

We see that for time modifications of this signal to occur they are constrained to variations in the amplitude A(t) and/or in the phase θ(t). Comparing the modulated signal of Equation 5-2 with the carrier signal of 5-1, it is seen that any changes to the carrier are impressed on the amplitude and/or the angle of the original signal c(t).

These observations reveal an important distinction in modulation types, that variations impressed upon the carrier are either in amplitude, in angle, or in both. While amplitude and angle modulations are used in both analog and digital modulations, the only modulation which uses both simultaneously (called quadrature amplitude modulation) will be covered in Chapter 6.

Before embarking on a discussion of the different modulation types, we should consider why modulation of a carrier by a message signal is required at all. When a message signal is generated–be it voice, music, data, etc.–it is generated at baseband. (For a discussion of baseband see Section 3.7.1.) Baseband signals can be thought of as low-pass signals, they dwell at the frequencies where they are generated. After creating a baseband message signal, ostensibly for communication with others at some distant location(s), what do we do with it? We probably cannot send this baseband to them directly (unless we are using directed propagation channels such as transmission lines or fiber optic cables). Instead, the message signal must be sent to them via a radio wave at a specified frequency or frequency band. For example, if the communications circuit is via High Frequency (HF), the signal will be sent within the frequency range of 3 - 30 MHz. Similarly, a VHF signal is in the range from 30 - 300 MHz, etc. In order to have the message signal travel within the constraints of the channel of the circuit (i.e., within the proper frequency band), we must somehow modulate a carrier frequency to carry the message.

A carrier frequency is a signal of a single frequency which falls within the frequency band of the channel through which we wish to communicate. There is no information or message on this signal, just a single frequency. In order to carry the message, the carrier must be modulated by the message signal, m(t).

At the receiver the message signal must be recovered from the modulated signal through demodulation. In this chapter we will introduce some ideal demodulators with which to extract m(t); while in Chapter xxx we will examine practical demodulators in order to evaluate their vulnerability to jamming. We begin this analog modulation study with amplitude modulation.

5.1 AMPLITUDE MODULATION

In amplitude modulation, the *amplitude* of the carrier signal is modified so that when it leaves the transmitter it will no longer have the constant amplitude term A_c given in Equation 5-1 as the sole determining factor of its amplitude. It will instead have an amplitude term which varies according to m(t). Since the modulated carrier will now contain the message signal, we will call this new signal s(t). There are several ways to modulate the amplitude of the carrier which we will now discuss. We begin with standard broadcast AM.

5.1.1 Amplitude Modulation with Carrier (AM)

5.1.1.1 Modulation

Let's define a modulated AM signal s(t) to be

$$s_{AM}(t) = c(t) + c(t)k_a m(t) \qquad \textbf{(5-3)}$$

where c(t) is the carrier signal defined in Equation 5-1. This modulated signal is seen to be the carrier added to the product of the carrier, the message signal m(t), and a constant k_a. This AM signal allows for ease of demodulation (as will be shown in Section 5.1.1.2) but notice the inefficiency of the power usage since the unmodulated carrier is contained in s(t) along with the term containing m(t). Observe that the message signal is modified by a sensitivity constant k_a which controls the maximum amplitude of the message signal with respect to the carrier amplitude. By combining terms we can see that

$$s_{AM}(t) = A_c\ [1 + k_a m(t)]\ \cos\ (2\pi f_c t). \qquad \textbf{(5-4)}$$

Knowing that the original amplitude of the carrier was A_c, we can readily see that the *amplitude* (or envelope) of the modulated signal s(t) is

$$a(t) = A_c\ (1 + k_a m(t)). \qquad \textbf{(5-5)}$$

We never want the quantity $k_a m(t)$ to be larger than unity, i.e.,

$$k_a\ m(t)_{\max} \le 1. \qquad \textbf{(5-6)}$$

If m(t) is a sinusoid (which for most message signals it will be a combination of sinusoids), then it will take on peak values of $\pm A_m$, the maximum value of m(t). For illustration, an AM signal where $c(t) = \cos(2\pi(10)t)$, $m(t) = \cos(2\pi t)$, and $k_a = 0.9$ is shown in Figure 5-1 below. The envelope tracing out a(t) is shown by the two solid lines. Notice that the envelope is representative of m(t).

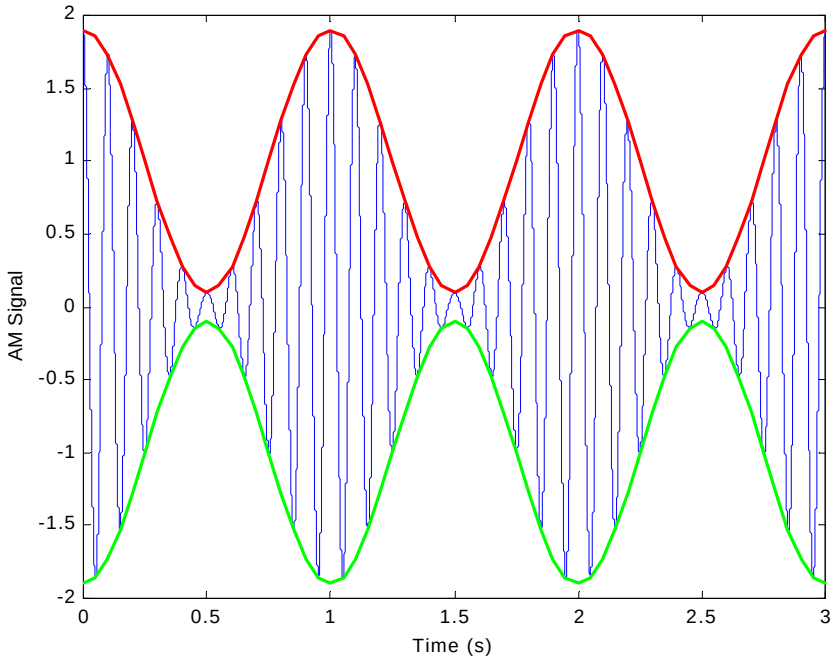

Figure 5-1. AM signal with $k_a A_m = 0.9$.

If the restriction of Equation 5-6 is not followed and $k_a m(t)_{max} > 1$, the value within the parentheses of Equation 5-5 will become negative (on the negative

alternation of m(t)). If this quantity is allowed to go negative, then a(t) will become negative; but a(t) is defined as the amplitude of s(t). A negative amplitude is not a defined quantity, as amplitudes are always positive. The result of a negative amplitude will be to cause *amplitude distortion* **as shown in the figure below, where c(t) and m(t) are the same as above, but k_a has been increased to 1.2. This signal is considered to have amplitude distortion because the envelope is no longer a faithful representation of m(t).**

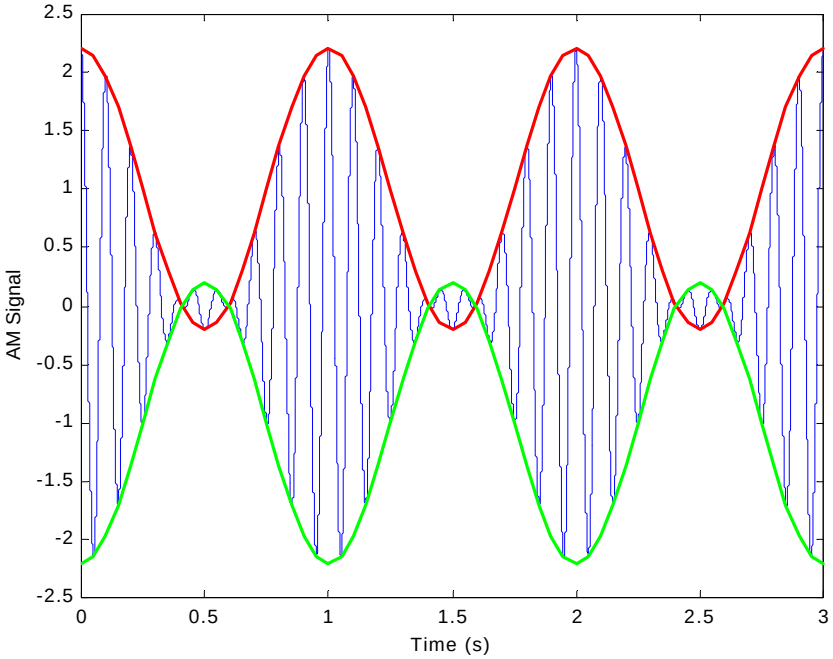

Figure 5-2. AM signal with $k_a\, m(t)_{max} = 1.2$.

From these two examples we see that to prevent amplitude distortion and maintain the envelope as a faithful representation of m(t) we must adhere to the requirement of Equation 5.6. This quantity $k_a \cdot m(t)_{max}$ is such a fundamental descriptor of the expected waveform that it is given its own name, *modulation index***, and symbol μ, so that**

$$\mu = k_a\, m(t)_{max} \leq 1. \qquad \textbf{(5-7)}$$

Since μ ≤ 1, its value multiplied by one hundred yields the percentage modulation,

$$\%\ modulation\ =\ k_a\ m(t)_{\max}\ (100)\ =\ 100\mu. \qquad \textbf{(5-8)}$$

so that a μ value of unity signifies 100% modulation.

Analysis of the AM equation of Equation 5-4 is facilitated by restricting m(t) to a single-frequency, single-amplitude signal. If the message signal is a sinusoid at frequency f_m with amplitude A_m, we will have

$$m(t)\ =\ A_m\ \cos\ (2\pi f_m t) \qquad \textbf{(5-9)}$$

and

$$\mu\ =\ k_a\ A_m. \qquad \textbf{(5-10)}$$

This allows us to write the AM signal modulated by a signal of a single frequency as

$$s_{AM}(t)\ =\ A_c\ [1\ +\ \mu\ \cos(2\pi f_m t)]\ \cos\ (2\pi f_c t). \qquad \textbf{(5-11)}$$

Multiplying this out we get

$$\begin{aligned} s_{AM}(t)\ &=\ A_c\ \cos(2\pi f_c t)\ +\ \mu\ \cos(2\pi f_m t)\ A_c\ \cos(2\pi f_c t) \\ &=\ A_c\ \cos(2\pi f_c t)\ +\ \frac{A_c \mu}{2}\Big[\cos[2\pi(f_c\ -\ f_m)t]\ +\ \cos[2\pi(f_c\ +\ f_m)t]\Big]. \end{aligned} \qquad \textbf{(5-12)}$$

This AM signal is seen to consist of the original carrier signal plus a signal at $f_c - f_m$ and another at $f_c + f_m$. The interpretation of this is that the original message signal (at frequency f_m) has been frequency shifted to two locations, one above the carrier and one below. These two signals each contain the information of the signal and are called *sidebands***. There are two sidebands with standard AM modulation, but the same information is contained in each. The amplitudes of these two additional signals are half the original carrier amplitude multiplied by the modulation index μ.**

To help clarify the outcome of Equation 5-12 we can find its Fourier transform. The Fourier transform will establish the spectral or frequency content of the signal. The F.T. of the first term is simple, from Equation 3-11 it is

$$A_c\ \cos\ (2\pi f_c t)\ \rightleftharpoons\ \frac{A_c}{2}\ [\delta(f\ -\ f_c)\ +\ \delta(f\ +\ f_c)]. \qquad \textbf{(5-13)}$$

The F.T. of the two right-side terms can be found similarly by letting $f_1 = f_c - f_m$ and $f_2 = f_c + f_m$ and again using Equation 3-11 to find

$$\frac{A_c\,\mu}{2}\cos(2\pi f_1 t) \rightleftharpoons \frac{A_c\,\mu}{4}[\delta(f - f_1) + \delta(f + f_1)]$$
$$= \frac{A_c\,\mu}{4}[\delta(f - f_c + f_m) + \delta(f + f_c - f_m)] \qquad \textbf{(5-14)}$$

and

$$\frac{A_c\,\mu}{2}\cos(2\pi f_2 t) \rightleftharpoons \frac{A_c\,\mu}{4}[\delta(f - f_2) + \delta(f + f_2)]$$
$$= \frac{A_c\,\mu}{4}[\delta(f - f_c - f_m) + \delta(f + f_c + f_m)]. \qquad \textbf{(5-15)}$$

Combining Equations 5-13 – 5-15 the F.T. of Equation 5-12 is found as

$$S(f) = \frac{A_c}{2}[\delta(f - f_c) + \delta(f + f_c)]$$
$$+ \frac{A_c\mu}{4}\left(\delta[f - (f_c - f_m)] + \delta[f + (f_c - f_m)]\right) \qquad \textbf{(5-16)}$$
$$+ \frac{A_c\mu}{4}\left(\delta[f - (f_c + f_m)] + \delta[f + (f_c + f_m)]\right)$$

which represents the carrier and the lower and upper sidebands.

Recall from Chapters 2 and 3 that a shifted delta function, e.g., $\delta(f - f_c)$, is just a delta function positioned at f_c. Therefore, Equation 5-16 is interpreted as, starting with the top line, delta functions at $\pm f_c$, and from the middle line, delta functions at $f_c - f_m$ and $-f_c + f_m$, and from the bottom line, delta functions at $f_c + f_m$ and $-f_c - f_m$. The carrier delta functions have a magnitude of $A_c/2$ and each of the sidebands has a magnitude of $A_c\,\mu/4$. The frequency domain representation of this AM signal is shown in Figure 5-3 where we have set $A_c = 2$, $A_m = 1$, and $\mu = 0.8$. (Frequency domain representation of m(t) included for comparison.)

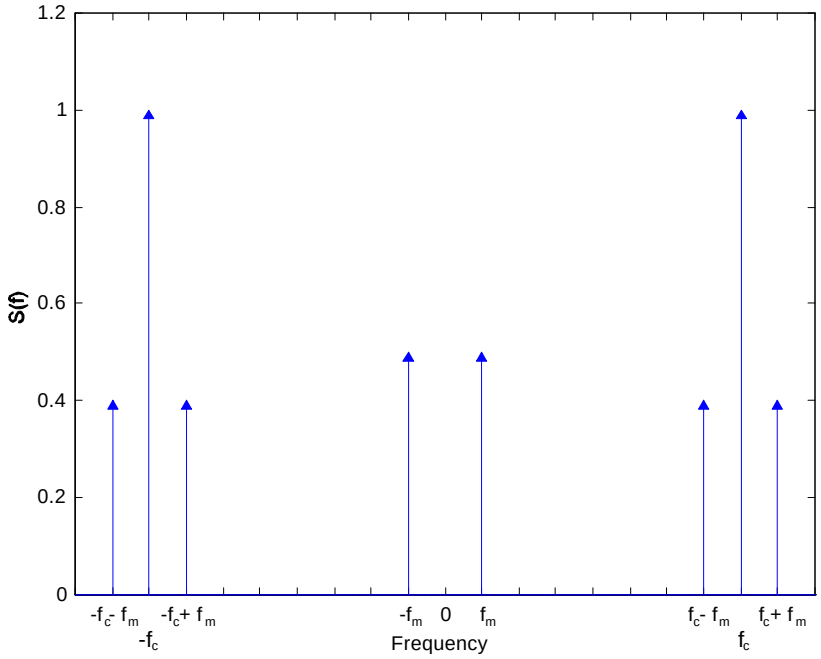

Figure 5-3. AM signal with $A_c = 2$, $A_m = 1$, and $\mu = 0.8$

Notice that the bandwidth of the modulating signal is f_m, but the bandwidth of the modulated signal is twice that or $2f_m$ because of the two sidebands around the carrier.

Using single-frequency modulating signals provides useful insight into the functioning of the AM signal, but usually a real-world modulating signal consists of many different frequencies. For analysis of a multi-frequency signal assume the frequency domain representation of the modulating signal is as represented in Figure 5-4. As seen in the figure M(f) is a baseband multi-frequency signal, extending from 0 Hz to f_{max}. Its bandwidth is therefore f_{max}.

To determine how the AM signal will appear in the frequency domain using this modulating signal, begin with Equation 5-2 and find its Fourier transform. The transform of the carrier is given by Equation 5-13. For the remainder of the equation, $A_c k_a m(t) \cos(2\pi f_c t)$, the transform can be found using Equation 3-21 to be

$$A_c k_a m(t)\cos(2\pi f_c t) \rightleftharpoons \frac{A_c k_a}{2}\left[M(f - f_c) + M(f + f_c)\right]. \qquad \textbf{(5-17)}$$

The F.T. is then the sum of Equations 5-13 and 5-17 as shown in Figure 5-5, where

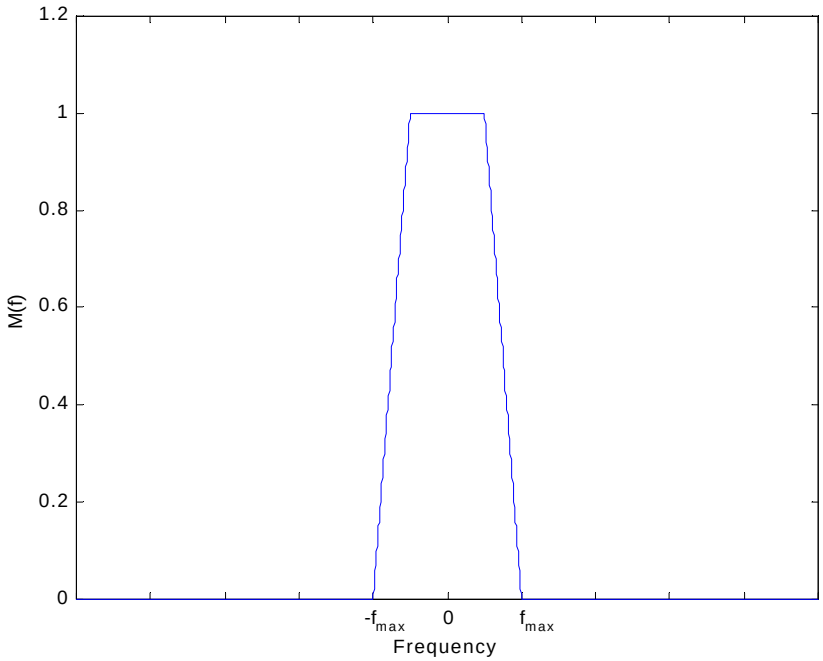

Figure 5-4. Frequency domain representation of multi-frequency m(t).

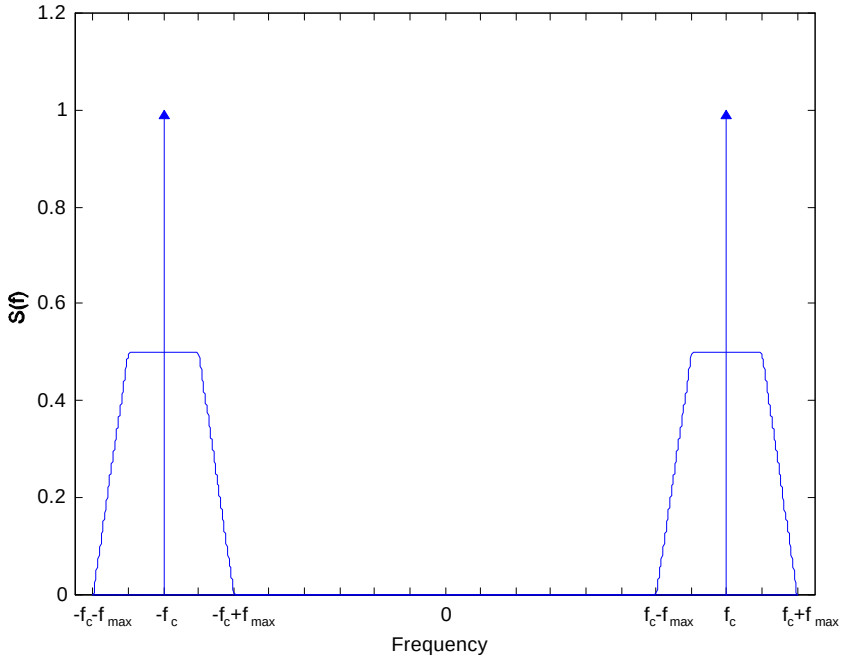

Figure 5-5. Frequency domain AM signal with modulating signal given in Figure 5-4, $A_c = 2$, and $\mu = 0.5$.

$A_c = 2$ and $\mu = 0.5$. Notice that the modulating signal has been shifted to $\pm f_c$ and that the bandwidth has doubled from f_{max} to $2f_{max.}$

Standard AM is characterized by this bandwidth doubling and that the carrier is transmitted along with the sidebands. Power transmitted in the carrier contains no information but allows for ease of demodulation. We will examine demodulation in Section 5.1.1.3 but first we will discuss the distribution of power in the AM signal.

5.1.1.2 Average Power in AM Signal

The average power of an AM signal can be found by integrating the square of the signal over one period and dividing by the length of the period, see Equation 2-29. Furthermore, the average power of *any* sinusoid of amplitude A, regardless of frequency, is $A^2/2$, as was shown in Equation 2-31. By inspection (i.e., without integrating), we see that the average power of Equation 5-12 is

$$\frac{A_c^2}{2} + \frac{A_c^2\mu^2}{4}\left(\frac{1}{2} + \frac{1}{2}\right). \qquad \textbf{(5-18)}$$

Therefore the power in the carrier is seen to be $A_c^2/2$ and that in the sidebands is $A_c^2\mu^2/4$.

The power being equally divided between the sidebands, each sideband then has a power content of $(A_c\mu)^2/8$. If $\mu = 1$, that is 100% modulation, the maximum power of the sidebands (i.e., where the information is) is only 1/3 of the total power of s(t). This is very inefficient communication.

5.1.1.3 Demodulation

When the incoming signal is collected at the receiver, the message information is not readily available from s(t). Just as the carrier was modulated with m(t) to get s(t), we must demodulate s(t) at the receiver to recover m(t). Demodulation is the process of isolating m(t) from all other time-varying signals. Isolating m(t) with the inclusion of multiplication by a constant is acceptable and usually the case.

Demodulation is also called detection, which is a holdover from earlier days with the crystal radio, where the crystal was a detector (diode) which performed the

demodulation. Two demodulator types are used prevalently for AM: the sometimes-used square law detector, and the envelope detector, which is used in nearly all AM receivers.

The square-law detector operates by taking the AM signal as input, squaring it, and passing the result to a low-pass filter. To see how it works, first square Equation 5-4 to get

$$\begin{aligned} s_{AM}^2(t) &= A_c^2\ [1 + k_a m(t)]^2\ \cos^2\ (2\pi f_c t) \\ &= A_c^2\ [1 + 2k_a m(t) + k_a^2 m^2(t)]\left[\tfrac{1}{2} + \tfrac{1}{2}\ \cos(4\pi f_c t)\right]. \end{aligned} \qquad \textbf{(5-19)}$$

After low-pass filtering, the term containing f_c will be eliminated leaving a DC term (which is easily removed) and the demodulated signal

$$s(t)_{demod} = \frac{A_c^2}{2}\ [2k_a m(t) + k_a^2 m^2(t)]. \qquad \textbf{(5-20)}$$

Notice that both m(t) and $m^2(t)$ outputs from the demodulator. The m(t) term has therefore not been completely isolated from all other time-varying signal. The $m^2(t)$ term represents distortion and can only be tolerated if $k_a \ll 1$ so that its contribution is negligible compared with that of m(t). With $k_a \ll 1$, the modulation index must be kept very low at the transmitter which reveals why this type demodulation is not often used.

The envelope detector is a simple circuit of a diode followed by a RC low-pass filter circuit, as shown in Figure 5-6 below. The operation of the circuit is simple. The diode only allows the positive half of $s_{AM}(t)$ to enter the filter. The positive voltage of s(t) rapidly charges the capacitor to its instantaneous value. As the wave goes negative the diode blocks the capacitor from discharging through the source and the capacitor must discharge slowly through the load resistor, maintaining the output voltage near its previous level. This continues until the next charging cycle from the input which will either increase the voltage on the capacitor or it will continue discharging.

Both these demodulation methods are known as non-coherent (as well as being non-linear). We will discuss coherent demodulation in Section 5.1.2.2.

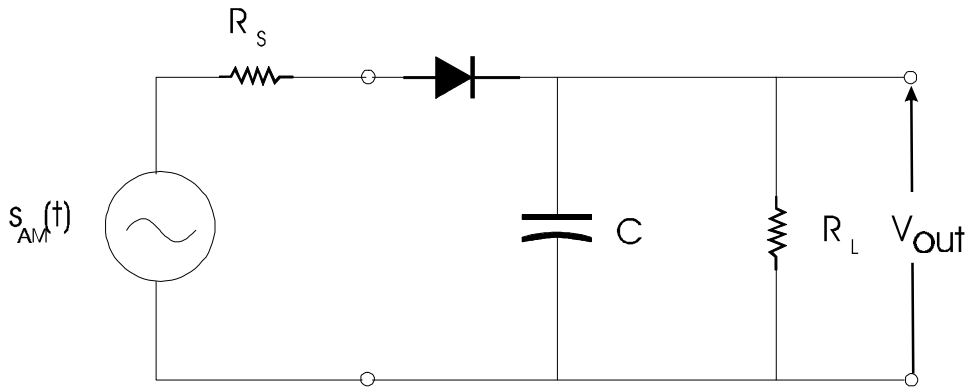

Figure 5-6. Envelope detector.

5.1.2 Double-Sideband Suppressed-Carrier Modulation

5.1.2.1 Modulation

In the last section we saw that the AM signal was defined as

$$s_{AM}(t) = c(t) + k_a\, m(t)\, c(t) = c(t) + s_m(t), \qquad \textbf{(5-21)}$$

where we define $s_m(t)$ to be the portion of the signal containing modulation. We also saw that the power in the sidebands (the information) was only a portion of the power transmitted in $s_{AM}(t)$. Specifically we can see that c(t) in the equation above consists of just the carrier. If we could somehow eliminate, or suppress, c(t) from $s_{AM}(t)$ and leave just $s_m(t)$, we could transmit the same information with much less power in the signal s(t).

This is exactly the idea behind double-sideband suppressed-carrier (DSBSC or just DSB) modulation. By suppressing the carrier we can see from the equation above that the DSB waveform will be of the form

$$s_{DSB}(t) = m(t)\, c(t) = A_c\, m(t)\, \cos(2\pi f_c t). \qquad \textbf{(5-22)}$$

The frequency domain DSB representation can be established by finding the Fourier transform of s(t). Using Equation 3-21 the F.T. is readily seen to be

$$S_{DSB}(f) = \frac{A_c}{2}\,[M(f - f_c) + M(f + f_c)]. \qquad \textbf{(5-23)}$$

If m(t) is a sinusoid of a single frequency, e.g., $A_m \cos(2\pi f_m t)$, then the modulated DSB signal will be

$$s_{DSB}(t) = A_c A_m \cos(2\pi f_m t) \cos(2\pi f_c t)$$
$$= \frac{A_c A_m}{2} \left[\cos\left(2\pi(f_c - f_m)t\right) + \cos\left(2\pi(f_c + f_m)t\right)\right]. \quad \textbf{(5-24)}$$

Notice that the two sidebands are still evident, just as with AM, but the isolated carrier signal is no longer present.

For c(t) = cos(2π(10)t) and m(t) = cos(2πt) (the same values given in the AM example in Figure 5-1), the DSB signal is as shown in Figure 5-7. The signal $s_{DSB}(t)$ is shown as the high frequency signal while m(t) is superimposed on top. Notice that the envelope of s(t) is no longer tracing m(t) as it did in standard AM, for example as in Figure 5-1. This indicates that the ubiquitous envelope detector used in AM will not demodulate DSB.

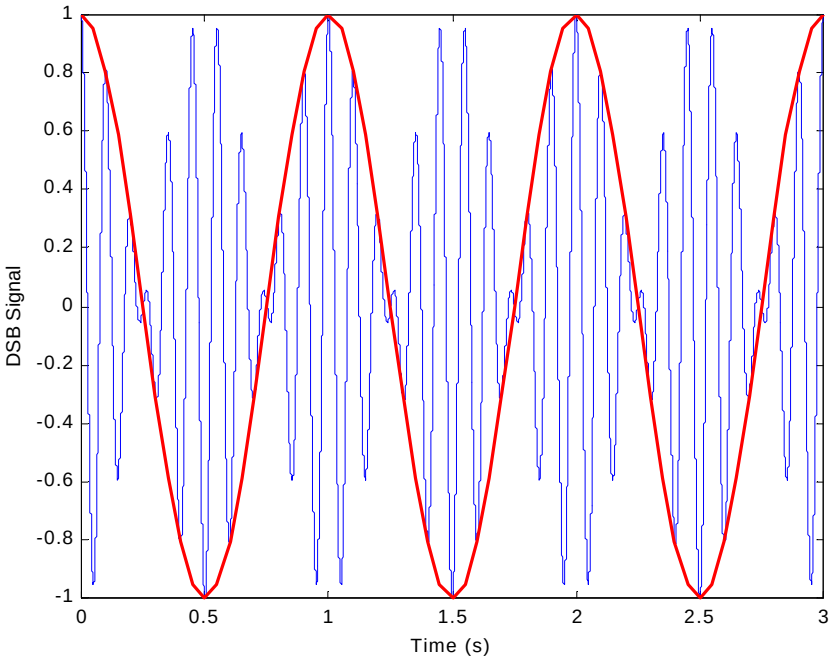

Figure 5-7. DSB signal with m(t) superimposed on top.

The frequency-domain representation for the signal of Figure 5-7 can be found using Equation 5-23 or 5-24 to be

$$S(f) = \frac{A_c A_m}{4}[\delta(f - (f_c - f_m)) + \delta(f + (f_c - f_m)) + \delta(f - (f_c + f_m)) + \delta(f + (f_c + f_m))] \quad \textbf{(5-25)}$$

Note that the frequency content of this signal is identical to that found for AM in Equation 5-16 and pictured in Figure 5-3 except that there is no carrier signal. There is still double sideband just as in AM so that the bandwidth required by the system is twice that of the message signal.

5.1.2.2 Demodulation

How do we demodulate this DSB signal? If we examine $s_{DSB}(t)$ we can see that squaring it yields $m^2(t)c^2(t)$; we cannot recover m(t) from this signal. What about envelope detection? Again examining $s_{DSB}(t)$ in Figure 5.7 we can see that the envelope of $s_{DSB}(t)$ is not the envelope of m(t). So this method will not work either. We need a new method to demodulate this signal.

Upon comparing the differences between DSB and AM, we find that DSB does not contain the carrier signal. It would seem that reinserting the suppressed carrier back into the DSB signal might facilitate demodulation. In order to insert the carrier back into the signal $s_{DSB}(t)$ we need to know and be able to synthesize the carrier frequency exactly. We will begin by allowing a constant phase difference between the original carrier c(t) and the synthesized carrier $c_s(t)$. Recall that the original is

$$c(t) = \cos(2\pi f_c t), \quad \textbf{(5-26)}$$

and we require the synthesized to be identical except for a possible phase offset,

$$c_s(t) = \cos(2\pi f_c t + \phi). \quad \textbf{(5-27)}$$

Let's multiply the signal $s_{DSB}(t)$ by this synthesized carrier in a product modulator and call the output v(t). We will get

$$v(t) = s_{DSB}(t)c_s(t) = A_c\ m(t)\ \cos(2\pi f_c t)\ \cos(2\pi f_c t + \phi)$$
$$= \frac{A_c\ m(t)}{2}\ [\cos(\phi) + \cos(4\pi f_c t + \phi)]. \qquad \textbf{(5-28)}$$

After low-pass filtering v(t) the output $v_o(t)$ will then be

$$v_o(t) = \frac{A_c}{2}\ \cos(\phi)\ m(t), \qquad \textbf{(5-29)}$$

which is the original message signal multiplied by a constant $\frac{1}{2}A_c\cos\phi$.

This process of multiplying the incoming signal by a replica of the carrier frequency then low-pass filtering the output is called *coherent* **or** *synchronous detection***. The envelope detector discussed above is an example of a** *noncoherent detector***.**

There are two problems with the output represented by Equation 5-29 and they both have to do with ϕ. First, what if ϕ is not constant but is a function of t, i.e., $\phi = \phi(t)$? (As you would have from a moving transmitter such as in an aircraft, see Chapter xxx). In this non-constant ϕ case there would be an independent time-varying signal cos $\phi(t)$ multiplied by m(t). This will prevent isolating m(t) without distortion.

Second, what if $\phi = \pi/2$? Now cos $\phi = 0$ and the product of Equation 5-29 leaves no message to detect. This situation is called the quadrature null effect.

The way to correct both these situations which prevent demodulation of m(t) is to force ϕ to zero. We will then have $c_s(t)$ identical to c(t) in frequency and phase. We do this with the Costas loop.

The Costas loop operates by first generating synthesized cosine and sine waves at the carrier frequency at some arbitrary phase ϕ with respect to c(t). These two signals are then used as the $c_s(t)$ frequencies to two coherent detectors identical to that above. The outputs from these detectors are $\frac{1}{2}A_c\cos\phi m(t)$ and $\frac{1}{2}A_c\sin\phi m(t)$. The $\cos\phi m(t)$ term is taken as the output as before while both are input to a phase discriminator. There they are multiplied together then low-pass filtered to get a

signal proportional to ϕ in magnitude and polarity. This is feedback to the $c_s(t)$ synthesizing oscillator which adjusts its output, forcing ϕ to zero.

5.1.3 Single-Sideband Modulation

We saw in both AM and DSB that the information signal m(t) was translated to the carrier frequency f_c. We observed that the signal bandwidth was doubled from that of the original bandwidth of m(t). This doubling was due to the positive frequencies of the baseband signal m(t) translating to above f_c while the negative frequencies of the baseband were translated to below f_c. All the information of m(t) is contained in the positive frequencies of the baseband (or alternatively in the negative frequencies), so that doubling the bandwidth is not necessary for information transfer. Doubling the bandwidth did, however, allow ease of generation and/or demodulation of the transmitted signal, s(t).

If we are willing to pay the price of complex modulation/demodulation of the signal, we can transmit just one sideband (i.e., single sideband, SSB) with no loss of information. In this way we can transmit less power and have a more efficient system, or boost the power and get further range, better SNR, etc.

5.1.3.1 Modulation

Just as we started analysis of DSB by comparing it to AM, we will begin our discussion of SSB by starting with DSB. Let's begin with the spectrum of the message signal, m(t). We might have a frequency domain representation something like Figure 5-4. Notice that the maximum value of this illustrated signal occurs at f = 0. The value here is M(f) evaluated at 0, or M(0). If we now multiply m(t) by the carrier, $c(t) = A_c \cos(2\pi f_c t)$, we translate M(f) (both positive and negative frequencies) out to $\pm f_c$ at a center amplitude of $\frac{1}{2}A_cM(0)$ as can be seen in Figure 5-5, which we know to represent DSB. If we now send the DSB signal through a filter whose bandwidth only allows the frequencies corresponding to the positive baseband signal through, we will have only the upper sideband as pictured in Figure 5-8. Similarly, if we only allow the lower sideband to pass, we will generate a signal similar to 5-9.

While filtering the unwanted sideband (called frequency discrimination) is one

method used to generate SSB, some precautions are mandated. Because real filters are not ideal, the lower sideband cannot simply be sliced from the upper sideband.

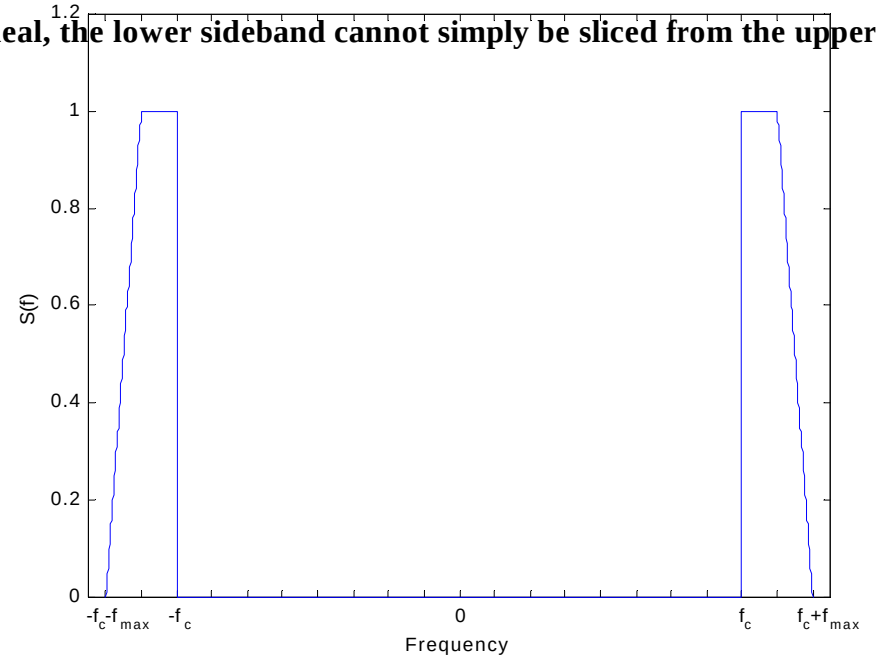

Figure 5-8. Single sideband signal consisting of only upper sideband.

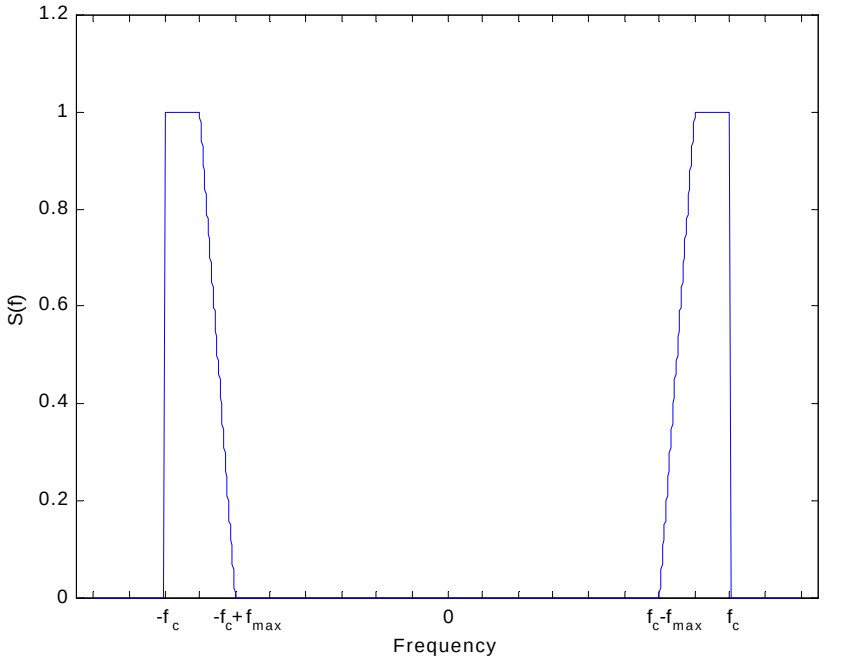

Figure 5-9. Single sideband signal consisting of only lower sideband.

Therefore, there must be some space between the positive and negative frequencies of the baseband m(t). This is usually the case for audio signals where the frequencies are not significant until around 300 Hz. For data signals however, the baseband might continue down to DC.

Another potential problem is that if the SSB signal is to be broadcast at radio frequency (RF), filtering the unwanted sideband at the RF may not be possible. In this case, a multi-stage translation may be required where the initial translation and filtering occur at a lower frequency followed by a translation to the RF and final filtering.

With the frequency domain of a SSB signal now established, we need the time domain representation of the signal. The mathematics to derive a time-domain single sideband signal are fairly lengthy, so will not be presented here. The interested reader should consult Appendix xxx. The result there is that an upper sideband signal can be represented by

$$s_{USB}(t) = \frac{A_c}{2}\,[m(t)\cos(2\pi f_c t) - \hat{m}(t)\sin(2\pi f_c t)], \qquad \textbf{(5-30)}$$

where $\hat{m}(t)$ is the quadrature of m(t), called the Hilbert transform of m(t). Equation 5-30 can be seen to be the original DSB signal minus its Hilbert transform. In a similar manner the lower sideband can be found as

$$s_{LSB}(t) = \frac{A_c}{2}\,[m(t)\cos(2\pi f_c t) + \hat{m}(t)\sin(2\pi f_c t)]. \qquad \textbf{(5-31)}$$

As verification that these two waveforms are indeed valid SSB representations, again let $m(t) = A_m \cos(2\pi f_m t)$. Then $\hat{m}(t) = A_m \sin(2\pi f_m t)$. Substituting into Equation 5-30 we have

$$\begin{aligned} s_{USB}(t) &= \frac{A_m A_c}{2}\,[\cos(2\pi f_m t)\cos(2\pi f_c t) - \sin(2\pi f_m t)\sin(2\pi f_c t)] \\ &= \frac{A_m A_c}{2}\cos[2\pi(f_c + f_m)t], \end{aligned} \qquad \textbf{(5-32)}$$

where we have used Equation Axxx in the last step. Observe that this signal contains only the upper sideband; the $f_c - f_m$ term has been eliminated. Similarly, a check of Equation 5-31 with these substitutions shows only the lower sideband present.

The method used for creating single sideband signals directly as indicated here, i.e., without filtering, is called phase discrimination rather than frequency discrimination. Examination of Equations 5-30 and 5-31 show that they are comprised of the original double sideband signal with its quadrature either added or subtracted. This results in phase cancellation, hence the name.

5.1.3.2 Demodulation

Since the SSB signal consists of a DSB term added to another term, we can use the same demodulation method that we used for DSB, i.e., coherent detection. With coherent detection we first multiply s(t) by the synthesized carrier, $c_s(t)$ to get v(t). This output is then passed through a low pass filter to get $v_o(t)$. Using this same method we used for DSB we get

$$\begin{aligned} v(t) = s_{SSB}(t)\, c_s(t) &= \frac{A_c}{2}\,[m(t)\cos(2\pi f_c t) \pm \hat{m}\sin(2\pi f_c t)]\cos(2\pi f_c t) \\ &= \frac{A_c}{4}\, m(t) + RF\ terms. \end{aligned} \qquad \textbf{(5-33)}$$

The output of the demodulator is therefore $v_o(t) = ¼\ A_c m(t)$.

With the presence of the quadrature component, SSB is not tolerant of a phase difference between c(t) and $c_s(t)$. If a phase difference ϕ exists, then the output from the demodulator will be

$$v_o(t) = \frac{A_c}{4}[m(t)\ \cos\phi + \hat{m}(t)\ \sin\phi], \qquad \textbf{(5-34)}$$

so that some method of keeping ϕ small, e.g., a costas loop, will have to be used with SSB just as with DSB.

5.2 ANGLE MODULATION

In amplitude modulation, we saw that we could impose the message signal upon the carrier signal by varying its amplitude. The phase and frequency of the carrier were left unchanged. In angle modulation the amplitude of the carrier is left unchanged but either the phase or frequency of the carrier is varied in a manner proportional to the message signal.

If we start with a carrier signal $A_c \cos(2\pi f_c t)$, the phase of the carrier waveform is $\theta(t) = \theta_c(t) = 2\pi f_c t$. If we now modify $\theta(t)$ linearly with a message signal, i.e., $\theta(t) = \theta_c(t) + k_p m(t)$ where k_p is the phase sensitivity (in radians per volt, radians per ampere, etc.) we will have a phase modulated waveform. A phase modulated waveform will therefore be

$$s_{PM}(t) = A_c \cos[\theta_c(t) + k_p m(t)] = A_c \cos[2\pi f_c t + k_p m(t)]. \quad \textbf{(5-35)}$$

A frequency modulated waveform takes a little different form. If we look at a modulated waveform at some time-varying instantaneous frequency f_i, we could characterize it as

$$s(t) = A \cos(2\pi f_i t) = A \cos(\theta(t)). \quad \textbf{(5-36)}$$

If the phase of this signal were known but not the frequency, we could determine the frequency by differentiating the phase to get

$$\frac{d\,\theta(t)}{dt} = 2\pi f_i \quad \Rightarrow \quad f_i = \frac{1}{2\pi}\frac{d\,\theta(t)}{dt}. \quad \textbf{(5-37)}$$

If we now require that the instantaneous frequency depend not only on just f_c, but also on the modulating signal, we get $f_i = f_c + k_f m(t)$. The constant k_f is the frequency sensitivity in hertz per volt (or hertz per ampere). Substituting for f_i in Equation 5-37

$$f_i = f_c + k_f m(t) = \frac{1}{2\pi}\frac{d\,\theta(t)}{dt}. \quad \textbf{(5-38)}$$

To find $\theta(t)$ we simply integrate the quantity $2\pi f_i$,

$$\theta(t) = 2\pi\int_0^t f_i dt = 2\pi\int_0^t [f_c + k_f\, m(t)]dt = 2\pi f_c t + 2\pi k_f \int_0^t m(t)dt. \qquad \textbf{(5-39)}$$

A frequency modulated waveform is therefore

$$s_{FM}(t) = A_c\cos\left(2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right). \qquad \textbf{(5-40)}$$

Comparing Equations 5-35 and 5-40 we see that they are of the same general form and that the frequency modulated waveform is simply a phase modulated waveform with the integral of m(t) modifying the phase rather than m(t) modifying it as in Equation 5-35. Similarly, a phase modulated waveform is just a frequency modulated waveform with the differentiated integral of m(t) modifying the waveform rather than the integral.

Given this relationship it suffices to analyze one or the other and the one not analyzed can be inferred from the other. Therefore we will only analyze Frequency Modulation (FM).

5.2.1 Modulation

In the last section we saw that the instantaneous frequency of a frequency modulated wave is

$$f_i = f_c + k_f m(t) \qquad \textbf{(5-41)}$$

so that the FM wave could be described by

$$s_{FM}(t) = A_c\cos\,(2\pi f_c t + 2\pi k_f \int_0^t m(t)dt). \qquad \textbf{(5-42)}$$

To begin analysis of this modulation form let's start with a single-frequency sinusoidal modulating signal of constant amplitude, i.e.,

$$m(t) = A_m\cos(2\pi f_m t). \qquad \textbf{(5-43)}$$

the instantaneous frequency of the FM wave will then be

$$f_i = f_c + k_f m(t) = f_c + k_f A_m \cos(2\pi f_m t). \quad \textbf{(5-44)}$$

It will necessary to know how much the instantaneous frequency varies or deviates from the carrier frequency. Inspecting Equation 5-44 it is seen that the maximum deviation will occur when the cosine term is maximum. Since the maximum amplitude of the cosine term is unity, this maximum deviation will be equal to $k_f A_m$. We therefore define this maximum frequency change as the frequency deviation Δf,

$$\Delta f = k_f A_m, \quad \textbf{(5-45)}$$

for the single-frequency modulating signal. The instantaneous frequency of the (single frequency) modulated signal can now be written as

$$f_i = f_c + \Delta f \cos(2\pi f_m t) \quad \textbf{(5-46)}$$

and the FM waveform as

$$\begin{aligned} s_{FM}(t) &= A_c \cos\left(2\pi f_c t + 2\pi\Delta f \int_0^t \cos(2\pi f_m t)dt\right) \\ &= A_c \cos\left(2\pi f_c t + \frac{\Delta f}{f_m}\sin(2\pi f_m t)\right). \end{aligned} \quad \textbf{(5-47)}$$

Just as we defined a modulation index μ for AM, we now define a modulation index for FM as $\Delta f / f_m$ and we call it β. Using this notation, the FM signal is defined as

$$s_{FM}(t) = A_c \cos(2\pi f_c t + \beta \sin(2\pi f_m t)). \quad \textbf{(5-48)}$$

We can find the in-phase and quadrature components (see Appendix xxx) $s_I(t)$ and $s_Q(t)$ of this signal using the trig identity Equation Axxx

$$s_{FM}(t) = A_c \cos[\beta \sin(2\pi f_m t)]\cos(2\pi f_c t) - A_c \sin[\beta \sin(2\pi f_m t)]\sin(2\pi f_c t), \quad \textbf{(5-49)}$$

so that $s_I(t) = A_c \cos(\beta \sin(2\pi f_m t))$ and $s_Q(t) = A_c \sin(\beta \sin(2\pi f_m t))$.

Recalling the techniques of the complex envelope (Appendix xxx), we know that

$$\tilde{s}(t) = s_I(t) + j s_Q(t). \quad \textbf{(5-50)}$$

Substituting our values for the in-phase and quadrature terms, we get

$$\begin{aligned}\tilde{s}_{FM}(t) &= A_c[\cos(\beta\sin(2\pi f_m t)) + j\sin(\beta\sin(2\pi f_m t))] \\ &= A_c e^{j\beta\sin(2\pi f_m t)},\end{aligned} \tag{5-51}$$

from Euler's identity.

It is clear that $\tilde{s}_{FM}(t)$ is periodic so that we can define it in a Fourier series,

$$\tilde{s}_{FM}(t) = \sum_{n=-\infty}^{\infty} c_n e^{j2\pi n f_m t} \tag{5-52}$$

where

$$\begin{aligned}c_n &= \frac{1}{T_0}\int_{-T_0/2}^{T_0/2} \tilde{s}(t)\, e^{-j2\pi nt/T_0}\, dt \\ &= f_m \int_{-1/2f_m}^{1/2f_m} \tilde{s}(t) e^{-j2\pi n f_m t} dt \\ &= f_m A_c \int_{-1/2f_m}^{1/2f_m} e^{j\beta\sin(2\pi f_m t) - j2\pi n f_m t} dt.\end{aligned} \tag{5-53}$$

Let $x = 2\pi f_m t$, then $dx = 2\pi f_m dt$ and the limits of integration will be $-\pi$ to π so that

$$c_n = \frac{A_c}{2\pi}\int_{-\pi}^{\pi} e^{j(\beta\sin x - nx)} dx = A_c J_n(\beta), \tag{5-54}$$

where $J_n(\beta)$ is the nth order Bessel function of the first kind with argument β (see Appendix xxx). This result allows us to rewrite the Equation 5-52 as

$$\tilde{s}_{FM}(t) = A_c \sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi n f_m t} \tag{5-55}$$

and recalling the relationship between the complex envelope, the pre-envelope, and s(t) from Appendix xxx, we know that $s_{FM}(t)$ can be found from the complex envelope of Equation 5-55 by shifting in frequency and taking the real part, i.e.,

$$s_{FM}(t) = A_c \, Re\left[\sum_{n=-\infty}^{\infty} J_n(\beta) e^{j2\pi(f_c+nf_m)t}\right]$$
$$= A_c \sum_{n=-\infty}^{\infty} J_n(\beta)\cos(2\pi(f_c + nf_m)t). \quad \textbf{(5-56)}$$

Upon inspection, we can see that this version of the FM signal is composed of a series of constants ($A_c J_n(\beta)$) multiplied by cosine waves of frequencies, f_c ($n = 0$), $f_c \pm f_m$ ($n = \pm 1$), $f_c \pm 2f_m$, etc., where the positive components make up the upper sidebands and the negative the lower sidebands. Theoretically, the sum consists of the infinite harmonics of f_m.

It is not obvious from their appearance but Equations 5-48 and 5-56 represent the same FM modulated signal. An observant reader will probably ask why we would go to the trouble to obtain 5-56 when 5-48 appears simpler and more manageable. The reason is that 5-48 offers no clue as to the frequency content of the FM signal. It is not intuitive that a single modulating frequency would result in an infinite number of harmonics as indicated by Equation 5-56.

We can define the FM waves created in this fashion as either narrow-band or wide-band. This definition arises from a comparison with AM. Recall that for tone modulation of an AM carrier we had the carrier and an upper sideband at $f_c + f_m$ and a lower sideband at $f_c - f_m$. If we define narrow-band FM to have this same characteristic, we can see that the infinite summation above is reduced to having Bessel function magnitudes non-zero only for $n = 0$ and $n = \pm 1$, i.e., $J_n(\beta) = 0$, for $|n| > 1$. We find this is true for $\beta \leq 0.3$, where $J_0(\beta) \approx 1$, and $J_1(\beta) \approx \beta/2$ (see Appendix xxx). Substituting these values into Equation 5-56 we find that the narrow-band FM signal is

$$s_{FM}(t) \approx A_c \cos(2\pi f_c t) + \frac{\beta A_c}{2}\cos[2\pi(f_c + f_m)t] - \frac{\beta A_c}{2}\cos[2\pi(f_c - f_m)t], \quad \textbf{(5-57)}$$

which is composed of the carrier and the two sidebands. This looks very much like the AM signal except that the lower sideband can be seen to be negative in this case.

For wide-band FM, β is not constrained and the sidebands consist of the infinite frequency harmonics of f_m. The magnitude of the components is controlled by the modulation index, β, and therefore $J_n(\beta)$.

How many of these sidebands are important for the transmission of the FM signal? Another way to state this is how much bandwidth do we require to adequately transmit the signal? We defined the frequency deviation Δf as the amount the modulated signal frequency varies from the carrier frequency. However, because of the infinite summation of the sidebands, the total frequency span will exceed Δf. Analysis of the Bessel function shows that its magnitude rapidly approaches zero for those sidebands above Δf. Therefore, the bandwidth of the modulated signal always exceeds Δf, but is limited.

In trying to define the bandwidth, J.R. Carson in the 1920s noticed that for large β, the bandwidth is approximately equal to $2\Delta f$. However, for small β the bandwidth is closer to $2f_m$, for a single tone, or 2W in general (where W is the highest frequency contained in the modulating signal). He proposed a bandwidth definition still used today called Carson's rule which is

$$B \approx 2\Delta f + 2W = 2(\Delta f + W). \qquad \textbf{(5-58)}$$

However, Carson's rule generally underestimates the bandwidth requirement of the signal. A better estimate is one called Carlson's rule which allows for more of the spectral lines. Carlson's rule is

$$B \approx 2(\Delta f + 2W). \qquad \textbf{(5-59)}$$

5.2.2 Demodulation

Now that we have modulated the frequency of the carrier, how do we recover the original message signal? Since the message is modulated within the signal, just as we did with AM, we must demodulate to recover the message, m(t). However, the methods of AM, i.e., envelope detection (without some pre-detection) and coherent demodulation, will not work with FM. FM demodulators are often called discriminators and work on different principles than we have seen before.

There are two primary methods of demodulation: direct and indirect. Beginning with direct demodulation we will discuss a method called frequency discrimination. To understand its operation, recall the equation for a FM wave,

$$s_{FM}(t) = A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right]. \qquad \textbf{(5-60)}$$

If we differentiate s(t) with respect to time we get

$$\begin{aligned} s'_{FM}(t) &= -A_c[2\pi f_c + 2\pi k_f m(t)] \sin\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right] \\ &= -2\pi f_c A_c \left[1 + \frac{k_f}{f_c} m(t)\right] \sin[...]. \end{aligned} \qquad \textbf{(5-61)}$$

We can see that the envelope of s′(t) is $A_1[1 + A_2 m(t)]$ which can be demodulated with an envelope detector. Therefore, to demodulate the FM signal we only need to differentiate it and send it through an envelope detector.

One may notice that the effective carrier frequency of Equation 5-60 above is not the same as f_c but deviates around it. As long as f_c is much greater than f_m the detector will not follow these changes in frequency and no distortion will occur.

The most prevalent form of indirect demodulation is the phase-locked loop (PLL). The PLL finds many uses in communications and electronics, and an understanding of its operation will be beneficial in later chapters.

Look first at the PLL shown in Figure 5-10

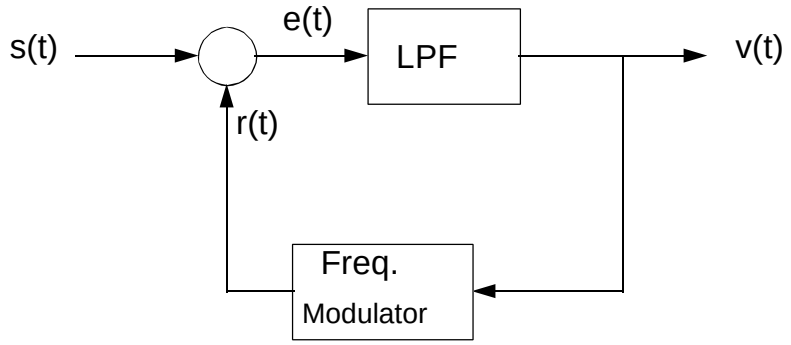

Figure 5-10. Phase-locked loop.

The input signal is s(t), the FM modulated wave

$$s(t) = A_c \cos\left[2\pi f_c t + 2\pi k_f \int_0^t m(t)dt\right] = A_c \cos(\theta_c + \theta_m). \quad \textbf{(5-62)}$$

The frequency modulator is a voltage controlled oscillator operating at the same frequency as the carrier but with a 90 degree phase shift, or $\sin(2\pi f_c t)$. The feedback signal r(t) is therefore a FM wave modulated by v(t)

$$r(t) = A_v \sin\left[2\pi f_c t + 2\pi k_v \int_0^t v(t)dt\right] = A_c \sin(\theta_c + \theta_v). \quad \textbf{(5-63)}$$

The multiplier has a gain of $-2k_m$ so that the error signal e(t) is

$$e(t) = -2k_m s(t) r(t). \quad \textbf{(5-64)}$$

Multiplying this out we see that

$$\begin{aligned} e(t) &= -2k_m A_c A_v \cos(\theta_c+\theta_m)\sin(\theta_c+\theta_v) \\ &= -2k_m \frac{A_c A_v}{2}[\sin(\theta_c+\theta_v-\theta_c-\theta_m) + \sin(\theta_c+\theta_v+\theta_c+\theta_m)], \end{aligned} \quad \textbf{(5-65)}$$

where we have used the trig identity

$$\cos(\alpha)\sin(\beta) = \frac{1}{2}(\sin(\alpha+\beta) + \sin(\alpha-\beta)). \quad \textbf{(5-66)}$$

Since the second sine term is at a frequency of $2f_c$, it will be removed by the low-pass filter. The remaining term will produce the output, v(t), of

$$v(t) = k_m A_c A_v \sin(\theta_m-\theta_v) = k_m A_c A_v \sin\theta_e, \quad \textbf{(5-67)}$$

where

$$\theta_m = 2\pi\ k_f \int_0^t m(t)\ dt$$

$$\theta_v = 2\pi\ k_f \int_0^t v(t)\ dt, \quad \textit{and} \qquad \textbf{(5-68)}$$

$$\theta_e = \theta_m - \theta_v.$$

After the PLL obtains lock, θ_e will be small so that $\sin\theta_e$ can be approximated by θ_e. We can now write the linear model of the PLL and

$$v(t) = k_m A_c A_v \theta_e = k_m A_c A_v \left[2\pi k_f \int_0^t m(t)dt - 2\pi k_v \int_0^t v(t)dt \right]. \qquad \textbf{(5-69)}$$

Differentiating both sides, we get

$$v'(t) = k_m A_c A_v [2\pi k_f m(t) - 2\pi k_v v(t)]. \qquad \textbf{(5-70)}$$

Transforming into the frequency domain we find that

$$V(f)(j2\pi f + k_m A_c A_v 2\pi k_v) = k_m A_c A_v [2\pi k_f M(f)] \qquad \textbf{(5-71)}$$

so that

$$V(f) = \frac{k_m A_c A_v k_f M(f)}{jf + k_m A_c A_v k_v}. \qquad \textbf{(5-72)}$$

In order for v(t) to be a true representation of m(t), the above equation must be a constant multiplied by the message signal. Since we have a varying quantity f, the only way for the equation to be a constant multiplied by M(f) is for $k_m A_c A_v k_v$ to be much larger than f, which is achieved through proper circuit design. With this constraint met, the output will be

$$v(t) = \frac{k_f}{k_v} m(t), \qquad k_m A_c A_v k_v \gg f. \qquad \textbf{(5-73)}$$